结合链路预测和ET机器学习的科研合作推荐方法研究*

吕伟民[1,2]　王小梅[3]　韩　涛[1]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)
[3](中国科学院科技战略咨询研究院　北京 100190)

摘要: 【目的】结合链路预测与机器学习，提出推荐未来科研合作的新方法，以提高单独基于链路预测方法的推荐精确度。【方法】构建加权作者合作网，以不同的链路预测指标作为特征输入，运用极端随机树(Extremely Randomized Trees, ET)机器学习算法训练分类，并利用遍历算法求取分类结果的最优权重组合，选取TOP准确度的预测作为合作推荐结果。【结果】选取纳米科技领域2008年–2010年SCI论文数据进行实证。在城市合作推荐中，改进的ET方法优于已有方法，有良好的推荐成功率；预测方法受网络结构等因素影响较小，适用范围更广泛。【局限】科研合作受合作动机、地域、语言等诸多因素影响，加权作者合作网没有反映在一篇论文中同城市、同机构的多个作者，也没有反映上述因素。【结论】改进算法能够比单个预测指标产生更准确的合作推荐建议，也为推广到大学等机构、个人等更微观的应用层面提供参考。

关键词: 科研合作网络　链路预测　机器学习　随机森林　极端随机树　推荐
分类号: G350

1　引　言

在知识经济时代，合作关系隐含着知识在某种社会关系之间的交流、转移、共享[1]。科研合作作为科学生产的一种重要形式，已成为科学研究成果增长和创新的强劲动力。从科研成果看，论文合著是科研合作最显性的表现之一，论文合著者之间的复杂关系构成了科研合作网络。科研合作网络会随时间推移而演化，学者们分别从网络结构[2-4]、网络演化机制[5]、网络增长[6-7]等方面研究科研合作网络，随后Kretschmer[8]又聚焦到个人合作行为的研究上。

近年来，复杂网络中链路预测方法受到越来越多的关注，链路预测在网络重构、网络演化模型评价、推荐系统[9-10]等方面有着重要的应用。因其在众多领域的重要影响及优越的预测性能，慢慢引进到图书情报领域。在该领域，针对科研合作网络的链路预测方法取得了很多进展[11-14]。但链路预测本身的预测精确度严重依赖网络拓扑结构，适用性较差。机器学习中的集成学习，通过将多种不同链路预测算法融合在一起，能够有效解决适用性较差这一局限，并在极大地扩展链路预测方法适用范围的同时，进一步提高链路预测方法的推荐准确度。

本文总结链路预测方法在科研合作网络中的研究进展；以纳米领域科研合作关系为例，基于链路预测和极端随机树，探讨推荐未来科研合作伙伴的方法；将链路预测与机器学习方法结合，对比分析Random Forest/Extremely Randomized Trees两种机器学习算法得出Extremely Randomized Trees算法预测精确度较

通讯作者：王小梅，ORCID：0000-0002-9895-1511，E-mail：wangxm@casisd.cn。
*本文系国家自然科学基金面上项目“科学结构特征及其演化动力学分析方法与应用研究”(项目编号：71173211)的研究成果之一。

好; 并进一步改进方法, 利用枚举得到的最优组合权重进行推荐排序, 得到更精确的推荐结果。为合作者本身和政策制定者产生精准合作推荐提供一种有效思路。

2 相关研究

网络中的链路预测是根据网络中节点的特征或已经存在的结构特征, 预测两点间边的存在性。Getoor 等[15]较早提出网络中的链路预测问题是指如何通过已知的网络节点以及网络结构等信息预测网络中尚未产生连边的两个节点之间产生连接的可能性。

链路预测在复杂网络领域有较深研究, Liben-Nowell 等[16]在链路预测的研究方法上做了开创性论述, 吕琳媛等[9, 17]将其引入国内, 总结了基于网络拓扑结构三种研究思路: 基于节点相似性的链路预测、基于最大似然估计的链路预测和基于概率模型的链路预测。三者各有优点与不足: 基于节点相似性的链路预测只涉及网络的结构信息, 相似性指标计算比较简单, 但不同指标在不同网络中的预测能力却不一致, 其预测的精确度高低取决于该种相似性能否很好地抓住目标网络的结构特征; 基于最大似然估计的链路预测由于针对的是整个网络结构, 计算复杂性较高, 不太适合在规模较大的网络中应用; 而基于概率模型的链路预测的优势在于较高的预测精确度, 不仅使用网络结构信息, 还涉及到节点属性信息, 但计算复杂度最高, 非普适性的参数使其应用范围受到限制。

因链路预测良好的性能, 因而受到来自不同领域、拥有不同背景的科学家的广泛关注。针对知识创造主体间的合作关系而构建合作网络, 以合著论文为基本表现形式。链路预测也被学者结合到科研合作网络推荐方法中, 相关研究大体分为 4 种方式(见表 1)。指标加权, 将合作的次数作为权重加入到指标中, 能在一定情况下提高链路预测精确度; 基于时序分析, 也是改进链路预测指标, 考虑时间因素来模拟演化过程; 不同层面网络对比, 分别从国家、机构、作者三个层面, 应用链路预测方法进行预测, 发现越宏观层面预测精确度越高; 加权网络, 从不同角度构造网络, 对每个网络分别应用链路预测方法, 将所得相似性分数加权后再进行排序。

表 1 链路预测在科研合作网络中的研究现状

主要方法	代表性研究
指标加权	Guns[18]以安德鲁大学学院合作网络以及计量情报学领域的合作网络为例, 得出加权的链路预测指标比不加权指标预测效果要好。
基于时序分析	Tylenda 等[19]考虑时间进化对预测结果的影响, 在 Wang 等[20]提出的局部概率模型基础上, 推导出考虑时间信息的最大熵原则方法, 把作者 a、b 最后一次合作到现在间隔的时间长度融入到加权的链路预测指标中, 提升链路预测的预测成功率。
不同层面网络对比	Yan 等[13]从作者、机构、国家三个层面构造合作网络进行研究, 对比三个层面合作网络在 8 种独立预测指标下的预测结果, 发现越高层面预测精确度越高, 即国家层面高于机构层面高于个人层面。
加权网络	Liben-Nowell 等[16]提出, 可以利用网络拓扑结构特征, 将论文标题、作者所在机构和地理位置信息加入到计算中, 对链路预测方法进行微调。具体实施时, Guns[21]将这些信息以不同层面的网络形式表现出来, 提出一种 Multi-Input 方法, 构建作者合作网络、部门网络和物理位置网络, 将三个子网络线性加权构成训练集。

这 4 类方法所采用的链路预测指标主要是基于节点相似性的链路预测指标, 但是要么直接针对单个链路预测指标分析, 要么对几个指标得到的相似性分数进行简单的线性加权, 这并不能达到很好的预测效果。至今为止, 已有超过 30 种指标被用在解决链路预测问题中, 但单一指标所考虑的信息都相对有限, 并且推荐成功率依赖网络本身的拓扑结构, 方法适用性较差。因此寻找一种合适的途径将这些指标集成, 或者是将更多的属性集成到一个指标中, 使模型具有更广泛的适用性以及较高推荐成功率, 成为学者研究的一个方向。

Mitchell[22]提出, 机器学习中的集成学习, 一个重要表现在于其结合各种特征, 取长补短利用多种形式的集成学习系统研究问题。集成学习也被尝试引入到科研合作领域。Guns 等[12]提出结合 RF 方法, 对非洲、中东和南亚的城市间科研合作进行研究, 构建了 1997

年–2001 年、2002 年–2006 年、2007 年–2011 年三个连续时间段内疟疾和肺结核研究领域中的加权合作网络。通过集成学习中的随机森林算法构建分类器，将预测排名靠前的结果作为科研合作建议进行推荐，推荐精确度优于单个链路预测指标的推荐精确度。

近年，机器学习越来越广泛地应用在各个学术领域，其与链路预测结合的思路已经开始有人关注[23]。本文改进 Guns 等[12]提出的结合 RF 的方法(以下统一将原方法称为"RF 方法"，改进的方法称为"改进 ET 方法")，将链路预测与机器学习结合，以提高科研合作推荐的准确性，使其更为实用。

3 数据说明与网络构建

本文进行纳米领域的科研合作推荐，利用 Arora 等[24]构建的检索式从 Web of Science(WOS)核心数据库中检索(检索式包含纳米领域各关键词，并去除部分无用的停用词，篇幅限制此处略去)，选取 2008 年、2009 年和 2010 年三个时期，每一年所有的纳米领域 Article 类型文章构成一个科研合作网络，三个时期形成三个持续变化的科研合作网络，如表 2 所示。

表 2 数据说明以及每个时期的 Article 论文数

数据说明	2008 年	2009 年	2010 年
论文数/篇	120 027	139 810	148 426
点个数	4 638	5 088	5 400
边条数	39 712	47 689	53 073

构建网络时，提取城市构建合作加权网，每一个城市代表网络中的一个点，所有的城市集合构成节点集。具体而言，一篇文章如果有两个作者，分别属于两个不同的城市，则这两个城市就被记录合作一次。同一文章如果有多个作者，多个作者属于不同城市，则这些城市在网络中存在一条连边，边的权重加 1。城市 A、B 间的权重为属于城市 A、B 的作者共著的文章数量。由于地址格式的很大差别以及数据的不一致，所有的结果都进行人工检查并进行必要的更正。

4 基于链路预测和 ET 的科研合作推荐

首先提取合适的链路预测指标特征，进而根据特征进行机器学习建模，然后在机器学习中的 ET 算法中融合不同的特征，从而取得更好的结果。

4.1 链路预测指标的特征选取

笔者关注目标国家中还没有合作的城市，对它们是否产生合作感兴趣。对于每个链路预测指标，依据现有的网络确定每个点对之间的相关性分数 *S*，挑选出那些还未连接的点对并依据 *S* 进行降序排序，可以产生一个未来最有可能合作的城市列表。

在三种链路预测研究思路中，基于节点相似性的链路预测方法计算简单，也可以尝试在较大规模网络中应用，同时在适当拓扑结构的网络中表现不错，因此在知识网络研究中应用较广。本文主要选取基于节点相似性的链路预测方法进行研究。综合考虑算法实施的效率以及预测表现[25]，选取 6 个指标作为机器学习的输入特征：考虑邻节点信息的指标，包括 Common Neighbours(CN), Adamic/Adar[26](AA), Resource Allocation(RA)；考虑整个网络拓扑结构的指标，包括 Katz[27], Graph Distance(GD), SimRank。Guns[18]的研究表明权重在指标中的恰当体现可以提高预测精确度，因此实验只包括加权的版本。

利用链路预测方法进行特征提取主要包括两个步骤[28]：将链路预测指标应用在某个训练网络上，预测可能产生的新链接；通过与测试网络进行比较，评价链路预测指标。

4.2 融合 ET 方法

随机森林(Random Forest, RF)和 ET 都属于机器学习中的集成学习，Random Forest 从特征集合中选择效果最佳的那个特征用来分类，得到的分类结果也许稍好些，但多次运行的结果可能不稳定。其后，ET 得以发展，完全随机地选择特征，得到的结果方差更小、更稳定。本文在对比两者后，主要选择 ET 作为训练方法，具体步骤如下。

①将数据集划分成 2008 年、2009 年和 2010 年三个时期，用 2008 年的数据构造早期网络 A1，2009 年的数据构造后期网络 A2，2010 年的数据作为验证集；

②选取相应链路预测指标，对 A1 中的每个点对，分别计算相关性分数；

③相关性分数作为特征，ET 根据 A1 中的特征以及 A2 中相对应的分类数据(是否连接)、作为训练集，A2 中已知的边是否存在(用 0，1 表示)作为分类结果进行学习，构建模型。通过学习当前时间片每个预测指标的相关性强度以及下一时间片对应是否产生连接，构建较为准确的分类器；

④对 A2 中每一个可能存在的连接，匹配 A1 中链路预

测指标得到的相关性分数;

⑤将 A2 中的特征作为训练集，利用之前步骤构造的分类器进行分类，提供预测的分类结果，挑选出重新判断的可能连接的点对;

⑥给步骤⑤中重新挑选的可能连接的点对赋予权重，每一个权重组合对应一组推荐精确度;

⑦枚举所有可能的权重组合，从中选取精确度最高的推荐结果对应的权重组合，推荐前 n 对预测结果作为合作配对推荐。

本文使用 Guns 提供的 LinkPred 中 Python 包来计算链路预测指标①，使用 Scikit-learn[29]进行机器学习训练，RF/ET 算法森林中树的棵数选择 400，推荐精确度均为 10 次计算结果的平均值。

4.3　评价指标

链路预测通常使用的评价指标有 AUC、Precision 和 Ranking Score，它们对预测精确度衡量的侧重点不同。其中 AUC 从整体上衡量算法精确度[30]，Precision 只考虑排在前 L 位的边是否预测准确[31]，而 Ranking Score 更多考虑所预测的边的排序[32]。

在实际应用中，决策者以及科研工作者感兴趣的是推荐有高潜力的合作，一般只关注最有可能合作的前几个合作团体，而不会关注几十名以后的合作团体。因此本文采用 Precision 来评价推荐结果，选取 Top10 的预测结果进行推荐。以 2008 年和 2009 年的网络作为训练集，为 2009 年以后的合作网络产生推荐，推荐得出 n 对得分最高的未连接的节点对，如果验证集在 2010 年存在，那么这个推荐就是成功的。定义 s 为成功推荐的个数，n 为推荐的总数。推荐的精确度定义为 $SR = {}^{s}/_{n}$，作为衡量推荐质量的一个指标。

5　结果分析

5.1　单个指标、RF、ET 精确度对比

依据相关性分数预测不同城市的研究所间的合作，分别对加权的 AA、CN、GD、Katz、RA、SimRank 6 个指标进行操作，预测结果精确度如表 3 和图 1 所示。其中，Weighted Katz : $\beta = 0.001$，Weighted Graph Distance : $\alpha = 1$，Weighted SimRank : $C = 0.8$。

表 3　城市层面推荐精确度

精确度	指标(Weighted)							
	AA	CN	GD	Katz	RA	SimRank	RF 方法	ET 方法
$n=5$	80%	80%	80%	60%	80%	60%	60%	60%
$n=10$	80%	80%	90%	80%	90%	40%	60%	80%
$n=20$	85%	80%	90%	80%	85%	30%	62%	80%

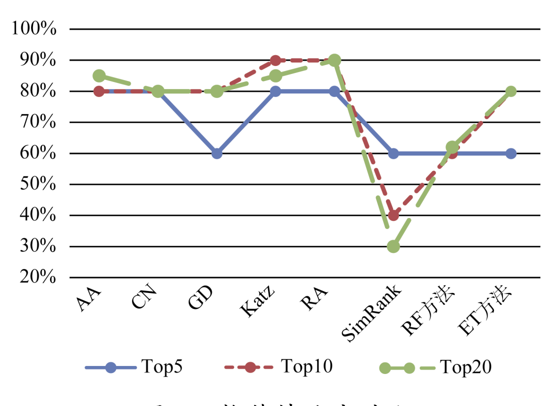


图 1　推荐精确度对比

已有的研究[12]结合 RF 方法，先利用链路预测指标作为训练特征，然后利用 RF 算法训练。使用 Gini Importance[33]作为权重确定每个预测指标对最终预测结果的相对贡献，即作为组合权重进行推荐排序。随机森林中每棵树的划分都记录指标异质性的降低。对森林中每个给定指标减少量的平均值就是该指标的 Gini Importance。简单而言，某链路预测指标 Gini Importance 越高，该指标越重要。

这里采用同样方法，同时利用 ET 算法进行训练，得到的推荐精确度结果见表 3。分别对比 6 个链路预测指标、RF 算法及 ET 的推荐精确度,可以看出：6 个链路预测指标，预测效果基本维持在 80%左右，除了 SimRank 指标的预测效果很差。通常情况下，预测精确度随着推荐个数的增加而下降，但个别指标也会存在不同的规律。出乎意料的是，已有的集成 RF 的方法计算得到的精确度只有 60%左右，这个精确度甚至比单个链路预测指标还要低。集成 ET 的方法，推荐精确度得到一定程度的提高，但是整体精确度依然不如某些单个链路预测指标的预测效果。因此本文对 Guns 的算法进行改进，一方面采用 ET 算法构建模型，另一方面采取遍历算法枚举所有可能的权重组合，从中选

①https://github.com/rafguns/linkpred/archive/stable.zip.

择预测效果最好的那对组合作为最优权重组合。

5.2 获取最优权重

在进一步改进的方法中，不直接采用 Gini Importance 作为权重，而是用遍历算法枚举所有可能的权重组合，每一组权重组合能够对应一组推荐精确度结果。为防止 ET 的预测结果存在过多偶然因素，精确度均为 10 次计算结果的平均值，选取精确度最好的前 5 组权重组合，如表 4 所示。

表 4　不同权重下改进 ET 的推荐精确度

Accuracy ([AA, CN, GD, Katz, RA])	$n=5$	$n=10$	$n=20$
[0.0, 0.0, 1.0, 0.0, 0.0]	100%	97%	85%
[0.05, 0.0, 0.85, 0.0, 0.1]	100%	90%	90%
[0.0, 0.05, 0.85, 0.0, 0.1]	100%	90%	90%
[0.0, 0.0, 0.9, 0.0, 0.1]	100%	90%	90%
[0.0, 0.0, 0.85, 0.05, 0.1]	96%	90%	90%

5.3 原因分析

改进 ET 方法的结果，预测精确度随推荐个数的增加而下降，但是在 Top10 的推荐结果中，推荐精确度等于或接近 100%，均优于单个链路预测指标，也优于 RF 和 ET 方法。原因如下。

(1) 直观上理解，如果说 5 个指标都认为某两个城市可能会产生合作，那么经过集成学习得到的结果就更倾向于产生合作。如果说只有两个指标预测出两个城市可能会产生合作，而其他三个指标均预测城市间不会产生合作，那么集成学习得到的结果可能就不会特别倾向于这两个城市产生合作。因此集成学习给出的推荐精确度要高于单个指标的推荐精确度。

(2) RF 方法中，融合了 6 个链路预测指标，这些指标本身的精确度有高有低，精确度低的那些指标会对最终预测结果产生影响。而改进的 ET 方法，是抽取了预测效果大于 70%的 5 个指标，可以理解成更良好的特征选择过程。从表 3 可知，SimRank 链路预测指标在本文数据集中预测效果很差，因此笔者在集成算法 ET 时，摒弃 SimRank 指标，只采用前 5 个指标进行集成。

(3) RF 方法预测精确度并不是很好，因为在有些情况下，Gini Importance 并不能十分准确反映权重，重要的指标不代表该指标的权重就越大，RF 方法在本文的数据集上应用效果并不是很好。改进 ET 方法因为遍历了所有可能的权重组合，更能得到最优精确度。同时，改进 ET 算法集成了不同特征的链路预测指标，可以有效针对不同拓扑结构的网络进行训练，适用范围更广泛。

改进 ET 算法稳定程度更高，给出的 5 组效果较好的权重组合，每个指标方差很小。比如 GD 指标，5 组组合中 GD 指标的权重区间为[0.85,1.0];RA 的权重区间也是[0.0,0.1]。因此可以放心地采用([AA, CN, GD, Katz, RA]:[0.0, 0.0, 1.0, 0.0, 0.0])作为最优权重组合。

在改进 ET 算法中，预测精确度较高的单个链路预测指标，其所对应的权重相对较高，说明在预测中起的作用较大。在实际应用中，可以先选择较多的链路预测指标进行预测，然后从中抽取效果较好的那些指标作为 ET 算法的特征输入，这样可以保证更好的机器学习效果。

与传统的图书情报领域单纯利用被引频次、有限因素等来构建网络进行链路预测相比，将链路预测方法与 ET 结合，参考更多因素来预测未来可能的合作关系，能够取得更好的方法适用性以及精确度。同时，这种机器学习方法还有一个突出的优势:对于处理大型合作网络问题能够很好地减少时间复杂度，提高预测效果。

6 结　语

本文介绍基于链路预测和 ET 改进的科研合作推荐方法，提出预测效果更好的 ET 算法，并在方法流程上加入枚举所有可能权重以求解最优化权重组合的步骤，选择最优化权重组合进行排序，极大提高了推荐准确度,提升后的推荐成功率高于所有单个指标的推荐成功率。ET 算法集成了不同特征的链路预测指标，可以有效针对不同拓扑结构的网络进行训练，使得推荐结果较稳定，适用性比单个链路预测适用性广，是一种很好的推荐研究合作伙伴的方法。同时相较于支持向量机等机器学习方法有较好的时间复杂度，在处理大型科研合作网络时有十分突出的优势，这使得科研合作推荐可以进行更微观的应用层面尝试。

本研究在构建数据集时，一篇文章如果有多位作者来自同一城市，按照一个作者来衡量。例如，一篇文章有 5 位作者来自 A 城市、3 位作者来自 B 城市，等同于这篇文章有一位作者来自 A 城市、一位作者来自

B 城市。后续研究中，将尝试从异质网的角度单独考虑同一城市多位作者的情况。

在城市层面，世界各国城市之间科研水平各有特点，寻找科研合作伙伴可以最大程度地发挥自己的科研优势。科研相对不发达的城市寻求与科研发达城市的合作也十分具有吸引力[34]，比如建立当地的精英中心以及对发展中国家共同的需求和问题有更全面的认识[35]。同时，本研究方法可以扩展到科研机构甚至科研工作者等更微观的层面，这样可以获取更实用的价值。但是微观层面构造的网络过于庞大，应用现有的工具进行链路预测分析效率很低，如果能改善现有链路预测方法或者改善数据集，得到各个链路预测指标的预测结果，就可以应用本文的方法进行合作推荐，这也是本文的一个研究目标。

整体而言，关于链路预测在科研合作推荐的应用还处在探索性、实证性的研究阶段，不同的预测指标和方法都有其优缺点与适用范围，因此寻找一种合适的途径来集成不同的预测结果或者是将尽量多的信息(如网络节点的属性、网络拓扑结构等)包含在指标中，是以后的研究方向。同时，借助链路预测的理论框架和评价方法，可以为交叉学科合作趋势、研发群体等提供推荐建议，从而解决寻找合作者这一难题[36]。更进一步的结论还需要更多实验支撑，这也是未来工作方向。

参考文献：

[1] 张斌，马费成. 科学知识网络中的链路预测研究述评[J]. 中国图书馆学报，2015, 41(3): 99-113. (Zhang Bin, Ma Feicheng. A Review on Link Prediction of Scientific Knowledge Network[J]. Journal of Library Science in China, 2015, 41(3): 99-113.)

[2] Newman M E J. Scientific Collaboration Networks. I. Network Construction and Fundamental Results[J]. Physical Review E, 2001, 64(1): 016131.

[3] Newman M E J. Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality[J]. Physical Review E, 2001, 64(1): 016132.

[4] Newman M E J. The Structure of Scientific Collaboration Networks[J]. Proceedings of the National Academy of Sciences, 2001, 98(2): 404-409.

[5] Barabási A L, Jeong H, Néda Z, et al. Evolution of the Social Network of Scientific Collaborations [J]. Physica A: Statistical Mechanics and Its Applications, 2002, 311(3-4): 590-614.

[6] De Solla Price D J. Little Science, Big Science⋯ and Beyond[M]. New York: Columbia University Press, 1986.

[7] Zuckerman H A. Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity[J]. American Journal of Sociology, 1968, 74(3): 276-291.

[8] Kretschmer H. Author Productivity and Geodesic Distance in Bibliographic Co-authorship Networks, and Visibility on the Web[J]. Scientometrics, 2004, 60(3): 409-420.

[9] Lü L, Zhou T. Link Prediction in Complex Networks: A Survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.

[10] Zhu B, Xia Y. An Information-theoretic Model for Link Prediction in Complex Networks[J]. Scientific Reports, 2015, 5: Article No. 13707.

[11] Guns R, Rousseau R. Predicting and Recommending Potential Research Collaborations [C]//Proceedings of ISSI. 2013: 1409-1418.

[12] Guns R, Rousseau R. Recommending Research Collaborations Using Link Prediction and Random Forest Classifiers[J]. Scientometrics, 2014, 101(2): 1461-1473.

[13] Yan E, Guns R. Predicting and Recommending Collaborations: An Author-, Institution-, and Country-level Analysis[J]. Journal of Informetrics, 2014, 8(2): 295-309.

[14] 张斌，李亚婷. 知识网络演化模型研究述评[J]. 中国图书馆学报，2016, 42(5): 85-101. (Zhang Bin, Li Yating. A Review of the Evolution Model of Scientific Knowledge Network [J]. Journal of Library Science in China, 2016, 42(5): 85-101.)

[15] Getoor L, Diehl C P. Link Mining: A Survey[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12.

[16] Liben-Nowell D, Kleinberg J. The Link Prediction Problem for Social Networks[J]. Journal of the Association for Information Science and Technology, 2007, 58(7): 1019-1031.

[17] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报，2010, 39(5): 651-661. (Lv Linyuan. Link Prediction on Complex Networks [J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.)

[18] Guns R. Missing Links: Predicting Interactions Based on a Multi-relational Network Structure with Applications in Informetrics [A]. // Missing Links: Predicting Interactions Based on a Multi-relational Network Structure with Applications in Informetrics[M]. Universiteit Antwerpen

(Belgium). 2012.
[19] Tylenda T, Angelova R, Bedathur S. Towards Time-aware Link Prediction in Evolving Social Networks[C]//Proceedings of the 3rd Workshop on Social Network Mining and Analysis. ACM, 2009: 1-10.
[20] Wang C, Satuluri V, Parthasarathy S. Local Probabilistic Models for Link Prediction[C]//Proceedings of the 7th IEEE International Conference on Data Mining. IEEE, 2007: 322-331.
[21] Guns R. Generalizing Link Prediction: Collaboration at the University of Antwerp as a Case Study[J]. Proceedings of the American Society for Information Science and Technology, 2009, 46(1): 1-15.
[22] Mitchell T M. Machine Learning. 1997 [J]. Burr Ridge, IL: McGraw Hill, 1997, 45(37): 870-877.
[23] Backstrom L, Leskovec J. Supervised Random Walks: Predicting and Recommending Links in Social Networks[C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. ACM, 2011: 635-644.
[24] Arora S K, Porter A L, Youtie J, et al. Capturing New Developments in an Emerging Technology: An Updated Search Strategy for Identifying Nanotechnology Research Outputs[J]. Scientometrics, 2013, 95(1): 351-370.
[25] Guns R. Bipartite Networks for Link Prediction: Can They Improve Prediction Performance[C]//Proceedings of ISSI. 2011: 249-260.
[26] Adamic L A, Adar E. Friends and Neighbors on the Web[J]. Social Networks, 2003, 25(3): 211-230.
[27] Katz L. A New Status Index Derived from Sociometric Analysis[J]. Psychometrika, 1953, 18(1): 39-43.
[28] Guns R. Link Prediction[A]//Measuring Scholarly Impact [M]. Springer International Publishing, 2014: 35-55.
[29] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2013, 12(10): 2825-2830.
[30] Hanley J A, McNeil B J. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve[J]. Radiology, 1982, 143(1): 29-36.
[31] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating Collaborative Filtering Recommender Systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53.
[32] Zhou T, Ren J, Medo M, et al. Bipartite Network Projection and Personal Recommendation[J]. Physical Review E, 2007, 76(2): 046115.
[33] Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees[M]. CRC Press, 1984.
[34] Schubert T, Sooryamoorthy R. Can the Centre–periphery Model Explain Patterns of International Scientific Collaboration Among Threshold and Industrialised Countries? The Case of South Africa and Germany[J]. Scientometrics, 2010, 83(1): 181-203.
[35] Boshoff N. South–South Research Collaboration of Countries in the Southern African Development Community (SADC)[J]. Scientometrics, 2010, 84(2): 481-503.
[36] Pavlov M, Ichise R. Finding Experts by Link Prediction in Co-authorship Networks[C]//Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics. 2007.

作者贡献声明:

吕伟民: 提出研究思路, 设计研究方案, 进行实验, 分析数据, 论文撰写;
王小梅: 论文最终版本修订;
韩涛: 数据分析, 方法分析。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: lvweimin@mail.las.ac.cn。

[1] 吕伟民. CITYCO_2008.txt. 2008 年纳米领域 article 文章, 以城市为维度表示合作.
[2] 吕伟民. CITYCO_2009.txt. 2009 年纳米领域 article 文章, 以城市为维度表示合作.
[3] 吕伟民. CITYCO_2010.txt. 2010 年纳米领域 article 文章, 以城市为维度表示合作.
[4] 吕伟民. CITYCO_2008.net. 2008 年纳米领域 article 文章, 提取城市作为科研合作网络的节点, 构建科研合作网络.
[5] 吕伟民. CITYCO_2009.net. 2009 年纳米领域 article 文章, 提取城市作为科研合作网络的节点, 构建科研合作网络.
[6] 吕伟民. CITYCO_2010.net. 2010 年纳米领域 article 文章, 提取城市作为科研合作网络的节点, 构建科研合作网络.

收稿日期: 2017-01-16
收修改稿日期: 2017-03-11

Recommending Scientific Research Collaborators with Link Prediction and Extremely Randomized Trees Algorithm

Lv Weimin [1, 2] Wang Xiaomei [3] Han Tao [1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper proposes a method to recommend scientific research collaborators based on link prediction and machine learning, which improves the precision of traditional method. [**Methods**] First, we used Link Prediction Algorithm index to build the feature input, and adopted the Extremely Randomized Trees Algorithm to train the classifier. Then, we obtained the optimal weight combination with the traversal algorithm to combine the classification results linearly. Finally, we received the best recommendation of collaborators. [**Results**] The improved ET method had better performance than the existing ones in recommending the collaboration cities. Besides, the proposed method was less affected by factors such as the network structure, and could be used with more applications. [**Limitations**] Scientific research collaboration is affected by the cooperation motivation, geographical, language and many other factors. The weighted author network did not examine authors from the same cities or with the same organizations. [**Conclusions**] The propsoed method could produce better recommendation results, which might help universities, institutions and individuals identify academic collabortors.

Keywords: Scientific Research Collaboration Network　Link Prediction　Machine Learning　Random Forest　Extremely Randomized Trees　Recommendation

---

OCLC Research 发布档案工作者与 IT 专业人士合作指南

OCLC 于近日发布了《Demystifying IT: 档案工作者与 IT 专业人士合作框架》，是"Demystifying Born Digital"系列的后续报告，该系列报告旨在帮助档案工作者更好地了解信息技术专业人员的工作方式，从而使其成为更加有效的合作者。

该报告的作者有克莱顿州立大学 Seth Shaw、密歇根大学图书馆 Richard C. Adler，和 OCLC Research 的 Jackie Dooley。这一报告描述了 IT 提供商的类型以及他们通常提供的服务，深入分析了软件开发过程，为建立伙伴关系提供指导，并强调资源约束的中心地位。

Dooley 说: "本报告旨在简要介绍信息技术，帮助数字档案管理员了解其特点，技术和文化，使其成为潜在的最有效的合作者。"

数字档案管理员需要工具和平台来提取、管理和提供所有类型的电子记录和数字内容的访问。数字系统的复杂性使得 IT 专业人员的参与变得至关重要。档案管理员具有复杂的领域知识，而 IT 人员具有先进的技术能力。有效合作需要了解彼此的专长，特点和制约因素。

(编译自: https://www.oclc.org/en/news/releases/2017/201711dublin.html)

(本刊讯)